		HQI+ResNet	66.1		COCO-(VQA) real images 1.0 multiple choice
14	Goyal et al. [15]	MCB	62.27	2017	VQA v2
		Deeper LSTM Q Norm	54.22		VQA v2
		LSTM Language Model only (blind model)	44.26		VQA v2
		Prior (most common answer in training set)	25.98		VQA v2
		MCB	62.3		COCO-(VQA) real images 2.0 open ended
		d-LSTM+nI	54.2		COCO-(VQA) real images 2.0 open-ended
15	Teney et al. [17]	Image features from bottom-up attention, adaptive K, ensemble	70.34	2017	VQA v2
16	Hu et al. [18]	N2NMN	64.2	2017	COCO-(VQA) real images 1.0 open-ended
17	Jiang et al. [20]	Pythia v0.1	70.24	2018	VQA v2
18	Anderson et al. [26]	Up-Down	70.3	2018	COCO-(VQA) real images 2.0 open-ended
19	Ruwa et al. [27]	AVQAN on SegAttendNet	54.7	2018	Visual7W
		AVQAN on original Visual7W	53.9		Visual7W

Table 1 shows the comparative study of several methods, datasets and performance issue of the previously used approaches for VQA. There are some of the research gaps in this area, as:

- The domain-specific datasets are required for effective VQA.
- Salient object detection from the image is still a challenging work for VQA.

- The semantic relationship is missing in image objects due to that relevant answers are uncertain in nature.
- Joint learning of visual question answering (VQA) and question relevance detection (QRD) is required to get efficient results.
- In present dataset questions and answers are biased in nature due to that efficiency of the present VQA methods and models are imperfect.
- Present VQA tasks answer confidently to the questions which are irrelevant to the given image.
- More precise and accurate results are an important requirement for VQA.

3. Rationale

Visual Question Answering is distinct from other traditional computer vision tasks such as object detection in an image; classify objects present in image etc. As these traditional tasks only extract the object present in the image, however, these tasks do not provide answers to the natural question related to the image content. The current state-of-the-art VQA systems are developed on the assumptions that the question asked will always be relevant to the image and therefore this system confidently provide incorrect answers to the irrelevant questions. However, before answering a question VQA systems should first check whether the asked is relevant to the image or not. To the best of our knowledge, we did not find any work which addressed this issue. Also, current VQA techniques are largely focused on the image processing tasks, however, less effort is made on the understanding the language semantics of the question.

The existing researches in VQA not much focus on the available semantic information present in the image. Also, Lu et al. [24] suggest that existing available datasets for VQA tasks are too much sparse and biased. These opportunities motivate us to pursue our research on the VQA task.

In our proposed work, we will attempt to utilize all the available semantic information present in the given image for the Visual Question Answering task. Our main objectives are as follows:

4. Objectives

1. To design algorithm for extraction and describing of image objects using natural language.
2. Extraction and identification of image semantics for Visual Question Answering.
3. Development of efficient Visual Question Answering mechanism.
4. Efficiency measurement of answers suitability for proposed Visual Question Answeringmechanism.

5. Research Methodology

The proposed methodology for the proposed work is as follows:

- Image data collection – In image data collection phase, we will use either real dataset or existing standard dataset like COCO, VQA v 2.0, etc.
- Identification of objects – In this phase, we will identify the different objects from the image.
- Identification of relation based on semantic information between the different objects – In this phase, we will calculate the semantic information between different objects from the image.
- Joint Knowledge – In this phase, we will establish the semantic relation between text and image object.
- Question and Answers –In question and answers phase, we will observe the gathered data to perform analysis.

References

[1]. Huijuan Xu, Kate Saenko, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering" arXiv: 1511.05234, 2015.

[2]. Bolei Zhou, Yuandong Tian, SainbayarSukhbaatar, Arthur Szlam, Rob Fergus, "Simple Baseline for Visual Question Answering" arXiv:1512.02167, 2015.

[3]. Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding" arXiv:1606.01847,2016.

[4]. IlijaIlievski, Shuicheng Yan, Jiashi Feng, "A Focused Dynamic Attention Model for Visual Question Answering" arXiv:1604.01485, 2016.

[5]. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh," VQA: Visual Question Answering" arXiv:1505.00468, 2016.

[6]. Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, Anthony Dick, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge" arXiv:1603.02814, 2016.

[7]. Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, Kate Saenko, "Modeling Relationships in Referential Expressions with Compositional Modular Networks" arXiv:1611.09978, 2016.

[8]. Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang, "Multimodal Residual Learning for Visual QA" arXiv:1606.01455,2016.

[9]. Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, "Stacked Attention Networks for Image Question Answering" arXiv:1511.02274, 2016.

[10]. Hyeonwoo Noh, Bohyung Han, "Training Recurrent Answering Units with Joint Loss Minimization for VQA" arXiv: 1606.03647, 2016.

[11]. Damien Teney, Qi Wu, Anton van den Hengel, "Visual Question Answering: A Tutorial," in *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 63-75, Nov.

2017.doi: 10.1109/MSP.2017.2739826

[12]. Damien Teney, Lingqiao Liu, Anton van den Hengel, "Graph-Structured Representations for Visual Question Answering" arXiv: 1609.05600, 2017.

[13]. Hedi Ben-Younes, RémiCadene, Matthieu Cord, Nicolas Thome, "MUTAN: Multimodal Tucker Fusion for Visual Question Answering" arXiv:1705.06676, 2017.

[14]. Jiasen Lu, Jianwei Yang, DhruvBatra, Devi Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering" arXiv:1606.00061, 2017.

[15]. Yash Goyal, TejasKhot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering" arXiv: 1612.00837. 2017.

[16]. Aishwarya Agrawal, AniruddhaKembhavi, Dhruv Batra, Devi Parikh, "C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset" arXiv: 1704.08243, 2017.

[17]. Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel, "Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge" arXiv:1708.02711, 2017.

[18]. Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Kate Saenko, "Learning to Reason: End-to-End Module Networks for Visual Question Answering " arXiv:1704.05526, 2017.

[19]. Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel, "FVQA: Fact-Based Visual Question Answering," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2413-2427, 1 Oct. 2018.doi: 10.1109/TPAMI.2017.2754246

[20]. Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, Devi Parikh, "Pythia v0.1: the Winning Entry to the VQA Challenge 2018" arXiv:1807.09956, 2018.

[21]. Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, Changyin Sun, "Inverse Visual Question Answering: A New Benchmark and VQA Diagnosis